Zero-Shot Learning via Contrastive Learning on Dual Knowledge Graphs

Jin Wang
Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University
Hefei, China
jinwang@stu.ahu.edu.cn

Bo Jiang*
Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University
Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
Hefei, China
jiangbo@ahu.edu.cn

Abstract

Graph Convolutional Networks (GCNs), which can integrate both explicit knowledge and implicit knowledge together, have shown effectively for zero-shot learning problems. Previous GCN-based methods generally leverage a single category (relationship) knowledge graph for zero-shot learning. However, in practical scenarios, multiple types of relationships among categories are usually available which can be represented as multiple knowledge graphs. To this end, we propose a novel dual knowledge graph contrastive learning framework to perform zero-shot learning. The proposed model fully exploits multiple relationships among different categories for zero-shot learning by employing graph convolutional representation and contrastive learning techniques. The main benefit of the proposed contrastive learning module is that it can effectively encourage the consistency of the category representations from different knowledge graphs while enhancing the discriminability of the generated category classifiers. We perform extensive experiments on several benchmark datasets and the experimental results show the superior performance of our approach.

1. Introduction

In recent years, zero-shot learning has attracted widespread attention in computer vision and machine learning areas. It aims to recognize the new categories that have never been appeared during the training process. The key of zero-shot learning is to leverage the knowledge from seen categories to describe unseen categories. In the past decade, researchers have proposed a large number of methods for zero-shot learning problems [4, 5, 8, 12, 20, 29]. For zero-shot tasks, it is usually necessary to explicitly explore the correlation relationships among different categories for knowledge transferring. To this end, Graph Convolutional Networks (GCNs) [15], which have powerful capabilities in exploiting category relationships, have been commonly employed for zero-shot learning tasks [12, 16, 19, 29]. Wang *et al*. [29] propose a GCN method to integrate implicit knowledge and explicit knowledge for zero-shot tasks. Kampffmeyer *et al*. [12] employ the dense graph propagation module to alleviate the over-smoothing problem in GCN for zero-shot classification. Liu *et al*. [19] propose a novel Attribute Propagation Network (APNet) to learn the classifier representation for each unseen category. The above methods [12, 16, 19, 29] explicitly exploit the relationships among different categories and demonstrate their effectiveness for zero-shot learning. However, the category relationships used in these models are relatively single type, which are unable to capture some more inherent category relationships (known as multiple knowledge graphs) in practical scenarios. To overcome this issue, recent works [3, 18] also explore multiple relationships among categories to perform zero-shot classification. For example, in work [3], a multi-relational GCN that integrates three category relationships (i.e., hierarchy, attribute and co-occurrence) is developed to generate the classifiers of unseen categories for zero-shot recognition tasks. Liu *et al*. [18] propose a novel Isometric Propagation Network (IPN) to combine the category representations from visual space and semantic space. However, previous methods [3, 18] generally employ an attention mechanism to perform the interaction and fusion for multiple knowledge graphs, which cannot fully exploit the inherent correlation and complementarity information across different knowledge graphs. Also, they usually lack of con-

*Corresponding author

sidering the discriminability of the generated classifiers of different categories which may lead to weak optimal performance for zero-shot learning.

Recently, contrastive learning has been demonstrated effectively for multi-view learning [10, 11, 30, 33]. Inspired by these works, we propose a novel zero-shot learning approach which aims to fully exploit *multiple knowledge graphs* for category representation and classifier prediction via contrastive learning. The proposed contrastive knowledge graph learning module can effectively encourage the consistency of the category representations obtained from different knowledge graphs while enhancing the discriminability of the generated category representations and classifiers. Overall, the main contributions of this paper are summarized as follows,

- We propose a novel dual knowledge graph contrastive learning approach to address the zero-shot classification tasks. The proposed approach can exploit multiple knowledge relationships among categories simultaneously to learn robust and discriminative classifiers for unseen categories.
- We introduce a graph contrastive learning scheme to address the general multiple knowledge graph representation and learning.

Experimental results on several benchmark datasets demonstrate the effectiveness of our method and show the benefits of the proposed dual graph contrastive learning module.

2. Related Works

2.1. Zero-shot Learning

Many of existing works [4, 8, 22, 24, 26, 32] are based on the implicit knowledge (i.e., semantic embeddings), which aim to learn a mapping function between semantic and visual representations. For example, Frome *et al*. [8] propose a novel DeViSE model that maps image features to semantic space for zero-shot classification. Norouzi *et al*. [22] leverage convex combination to achieve the mapping from visual feature to semantic feature space. In addition, explicit knowledge (i.e., knowledge graph) has also been employed for zero-shot learning, which aims to directly model the relationships among different categories [5, 25]. For example, Deng *et al*. [5] employ Hierarchy and Exclusion (HEX) graphs to capture the relationships among objects. Salakhutdinov *et al*. [25] adopt a novel hierarchical classification model, which effectively performs knowledge transfer from seen classes to unseen classes.

Recently, Graph Convolutional Networks (GCNs) based approaches that integrate both implicit knowledge and explicit knowledge have shown excellent performance on zero-shot learning tasks [12, 16, 19, 29]. As demonstrated in works [12, 29], with the semantic embedding and category hierarchical relationships, the visual classifiers of unseen categories can be effectively learned for zero-shot classification. Liu *et al*. [19] propose an Attribute Propagation Network (APNet) to generate unseen category representations for zero-shot classification. Li *et al*. [16] use a graph neural network to model the structural relationships among different categories and obtain better visual features of unseen categories. Recent works [3, 18] also exploit multiple knowledge graphs on zero-shot learning tasks. For example, in work [3], a multi-relational GCN model that combines three type category relationships is employed to generate the unseen class classifiers. Liu *et al*. [18] propose Isometric Propagation Network (IPN) to fuse category representations in both semantic and visual spaces.

2.2. Contrastive Learning

As a kind of self-supervised learning method, contrastive learning has been widely employed in multiple view learning tasks, such as node classification [27, 33], graph classification [10, 30] and few-shot learning [13, 17] etc. For example, Zhu *et al*. [33] adopt a novel GCA model for graph node classification by enhancing the consistency of node representations among different graph views. Velickovic *et al*. [27] propose Deep Graph Infomax (DGI) to learn robust graph node representations by maximizing the mutual information between the local and global graph representations for node classification task. Hassani *et al*. [10] propose to contrast local and global representations across different views for graph classification problems. Recently, some contrastive learning methods [9, 28] have also been employed for zero-shot learning tasks. Han *et al*. [9] propose a generalized zero-shot learning framework that uses contrastive embedding model to generate more discriminative visual samples. Wang *et al*. [28] employ Dual-Contrastive Embedding Network (DCEN) to learn more discriminative image feature representations for zero-shot learning tasks.

3. The Proposed Approach

Problem Formulation: In zero-shot learning, there are a total of n categories, which contain T seen/training categories (denoted as $\mathcal{Y}^T = \{y_1, y_2 \cdots y_T\}$) and U unseen categories (denoted as $\mathcal{Y}^U = \{y_{T+1}, y_{T+2} \cdots y_{T+U}\}$), i.e., $n = T + U$. There is no overlap between training seen categories $\mathcal{Y}^T$ and unseen categories $\mathcal{Y}^U$, i.e., $\mathcal{Y}^T \cap \mathcal{Y}^U = \emptyset$. The training categories include $\mathcal{V}$ labeled images $\mathcal{D} = \{(x_i, y_i)\}_i^{\mathcal{V}}$, where x_i denotes the i-th image instance and $y_i \in \mathcal{Y}^T$ represents the corresponding category. Note that, there are no labeled images for the unseen categories $\mathcal{Y}^U$. Meanwhile, let $S \in \mathbb{R}^{n \times p}$ be the semantic embedding vectors for all n categories, where p denotes the dimension of semantic vectors. The goal of zero-shot learning is to predict classifiers $\widetilde{W}^U \in \mathbb{R}^{U \times d}$ of unseen categories based on

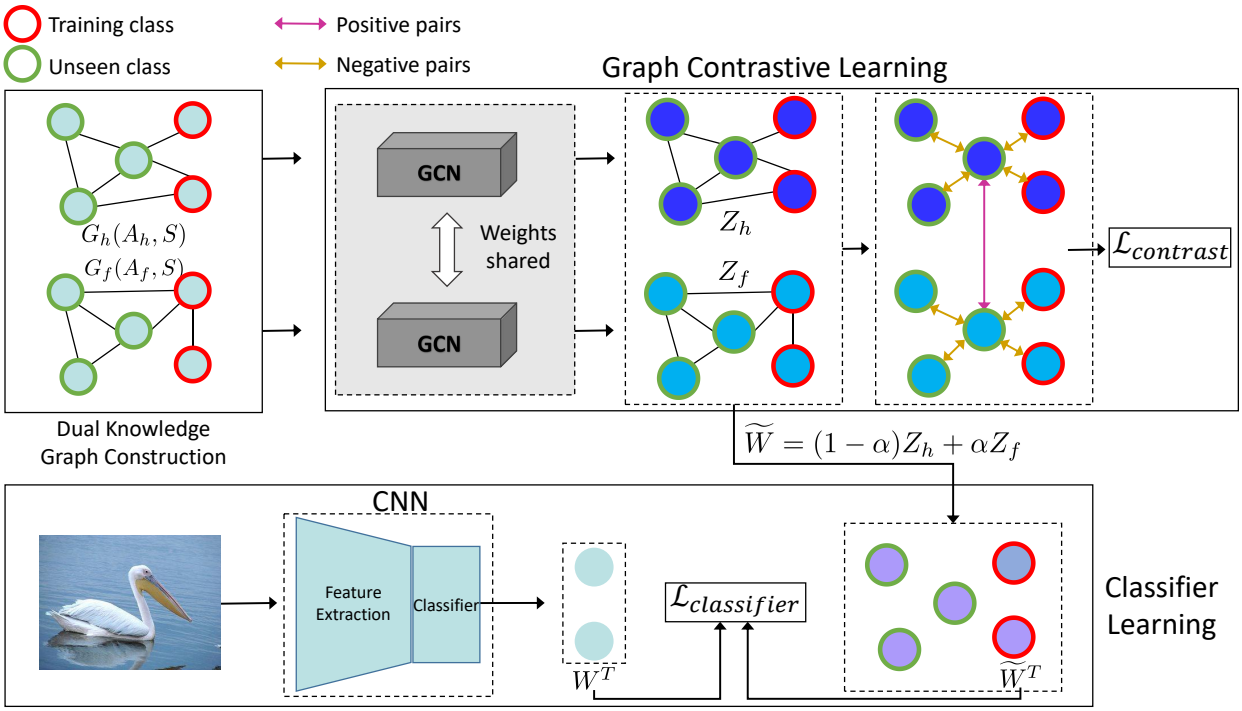

Figure 1. Overall framework of the proposed zero-shot recognition model. It is composed of three main parts, i.e., dual knowledge graph construction, graph contrastive learning and classifier learning.

the seen categories and perform classification for the testing unseen images, where d denotes the dimension of classifier weight vectors.

3.1. Overview

The overall architecture of our model is shown in Figure 1. Given two kinds of knowledge graphs as inputs whose nodes refer to specific categories, our goal is to learn effective visual classifiers for the unseen classes and then perform zero-shot classification, as suggested in works [12, 29]. The proposed zero-shot framework contains three main modules, i.e., dual knowledge graph construction, graph contrastive learning and classifier learning.

- **Dual knowledge graph construction.** We construct two knowledge graphs from different perspectives to capture more inherent relationship information among different categories.
- **Graph contrastive learning.** The purpose of graph contrastive learning module is to enforce the representations of corresponding nodes from different graphs to be consistent and the feature representations of different nodes within the same graph to be discriminative.
- **Classifier learning.** We obtain the final visual classifiers by fusing the graph node embeddings from two different knowledge graphs, and further employ the error loss function to learn the classifiers in a semi-supervised manner.

3.2. Dual Knowledge Graph Construction

Hierarchy knowledge graph. In order to exploit the hierarchical structure relationships among different categories, we adopt WordNet [21] hierarchy graph $G_h(A_h, S)$, as used in works [12,29], which is constructed based on expert knowledge. Here, $S \in \mathbb{R}^{n\times p}$ represents the semantic embedding representations of all n categories, where n, p denote the category number and semantic feature dimension, respectively. $A_h \in \mathbb{R}^{n\times n}$ encodes the hierarchical relationships among different categories.

Semantic correlation knowledge graph. In addition to the above hierarchy knowledge graph, we further explore the semantic correlations among different categories in the semantic feature space and construct a semantic correlation knowledge graph (K-nearest neighbor graph) $G_f(A_f, S)$ based on semantic embeddings S. Here, $A_f \in \mathbb{R}^{n\times n}$ represents the structural relationships among different categories in the semantic correlation knowledge graph G_f. Specifically, the adjacency matrix $A_f \in \mathbb{R}^{n\times n}$ is defined as

$$A_f(i,j) = \begin{cases} 1 & \text{if } j \in \mathcal{N}_K(i) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\mathcal{N}_K(i)$ denotes the top K nearest neighbors of node i in the semantic feature space.

3.3. Graph Contrastive Learning

3.3.1 Graph convolutional module

Similar to previous works [12, 29], we employ Graph Convolutional Networks (GCNs) [15] to conduct knowledge graph representation and learning. Specifically, given hierarchical graph $G_h(A_h, S)$ containing n nodes, we perform graph convolution propagation as follows

$$H_h^{(l+1)} = \sigma\big(D_h^{-1}\widehat{A}_h H_h^{(l)}\Theta^{(l)}\big) \tag{2}$$

where $l = 0, 1 \cdots L-1$ and L represents the number of convolutional layers. $H_h^{(0)} = S$ denotes the input node features and $H_h^{(l)}$ denotes the output feature representations of the l-th hidden layer. $\widehat{A}_h = A_h + I$ and D_h is the diagonal matrix with $D_h(i,i) = \sum_j \widehat{A}_h(i,j)$ and I is the identity matrix. $\Theta = \{\Theta^{(0)}, \Theta^{(1)} \cdots \Theta^{(L-1)}\}$ refer to the trainable weight matrices of GCN model.

In addition, for semantic correlation knowledge graph $G_f(A_f, S)$, let $H_f^{(0)} = S$ as input. Then, we employ the same graph convolution paradigm as in hierarchy graph G_h and conduct layer-wise propagation as

$$H_f^{(l+1)} = \sigma\big(D_f^{-1}\widehat{A}_f H_f^{(l)}\Theta^{(l)}\big) \tag{3}$$

where $l = 0, 1 \cdots L-1$ and σ refers to the Leaky ReLU activation function. $\widehat{A}_f = A_f + I$ and $D_f(i,i) = \sum_j \widehat{A}_f(i,j)$. $\Theta = \{\Theta^{(0)}, \Theta^{(1)} \cdots \Theta^{(L-1)}\}$ are the trainable weight matrices of the GCN model. Note that, the weight parameter matrix set Θ for both G_h and G_f are shared, as suggested in works [1, 7].

3.3.2 Contrastive learning

Let $Z_h = H_h^{(L)}$ and $Z_f = H_f^{(L)}$ be the outputs of the above two GCN branches. When the graph node representations $Z_h = \{z_h^{(1)}, z_h^{(2)} \cdots z_h^{(n)}\}$ and $Z_f = \{z_f^{(1)}, z_f^{(2)} \cdots z_f^{(n)}\}$ are obtained from knowledge graphs G_h and G_f respectively, inspired by works [10, 27, 33], we can use the contrastive learning module to constrain Z_h and Z_f. Specifically, the contrastive module encourages the representations of corresponding nodes in Z_h and Z_f to be more consistent and the representations of different nodes within knowledge graph node representations Z_h or Z_f to be more discriminative.

Formally, each node embedding representation $z_h^{(i)}$ generated from graph G_h can be regarded as an anchor, embedding $z_f^{(i)}$ generated from graph G_f forms the positive sample, and the other m node embeddings obtained by random sampling in graph G_h are treated as the negative samples. Then, for each positive pair $(z_h^{(i)}, z_f^{(i)})$, following the similar strategy [27], the objective function of contrastive learning is formulated as

$$\mathcal{L}(z_h^{(i)}, z_f^{(i)}) = -\Big(\log\varphi(z_h^{(i)}, z_f^{(i)}) + \sum_{j\neq i}^{m}\log\big(1-\varphi(z_h^{(i)}, z_h^{(j)})\big)\Big) \tag{4}$$

where $\varphi(z_h^{(i)}, z_f^{(i)}) = \sigma(z_h^{(i)} \cdot z_f^{(i)})$ denotes the critic function and σ refers to the sigmoid nonlinearity function. Since both knowledge graphs G_h and G_f are symmetric, similar to positive pair $(z_h^{(i)}, z_f^{(i)})$, we can also define the loss function on positive pair $(z_f^{(i)}, z_h^{(i)})$ as follows,

$$\mathcal{L}(z_f^{(i)}, z_h^{(i)}) = -\Big(\log\varphi(z_f^{(i)}, z_h^{(i)}) + \sum_{j\neq i}^{m}\log\big(1-\varphi(z_f^{(i)}, z_f^{(j)})\big)\Big) \tag{5}$$

Finally, the overall loss function of the proposed graph contrastive learning model is defined as

$$\mathcal{L}_{contrast} = \frac{1}{2n}\sum_{i=1}^{n}\Big(\mathcal{L}(z_h^{(i)}, z_f^{(i)}) + \mathcal{L}(z_f^{(i)}, z_h^{(i)})\Big) \tag{6}$$

3.4. Classifier Learning for Zero-shot Classification

In this section, we present how to obtain a visual classifier for each unseen category and employ the learned classifiers to perform zero-shot classification, as shown in Figure 1. Specifically, we first obtain the visual classifiers $\widetilde{W} \in \mathbb{R}^{n\times d}$ for all n categories by fusing both the graph node embeddings Z_h and Z_f together as

$$\widetilde{W} = (1-\alpha)Z_h + \alpha Z_f \tag{7}$$

where α is the weight parameter. Note that $\widetilde{W}$ consists of both T training and U unseen category classifiers, i.e., $\widetilde{W} = \{\widetilde{W}^T, \widetilde{W}^U\}$. Then, inspired by works [12, 29], the loss function based on T training categories can be formulated as

$$\mathcal{L}_{classifier} = \frac{1}{2T}\sum_{i=1}^{T}\sum_{j=1}^{d}\big(W_{ij}^T - \widetilde{W}_{ij}^T\big)^2 \tag{8}$$

where $\widetilde{W}^T \in \mathbb{R}^{T\times d}$ denotes the learned visual classifiers of T training categories. We can extract the last layer weight matrix of the pre-trained CNN as the ground truth classifier weights $W^T \in \mathbb{R}^{T\times d}$ for the T training categories, as suggested in previous works [12, 29]. Finally, our model contains graph contrastive learning and classifier learning modules. Therefore, the total loss is

$$\mathcal{L}_{all} = \mathcal{L}_{classifier} + \lambda\mathcal{L}_{contrast} \tag{9}$$

where λ is the weight parameter.

Inference Process. As shown in Figure 1, for a testing image from the unseen categories, we first use the pre-trained CNN to extract the image feature $x_t \in \mathbb{R}^d$. Then,

we employ the learned unseen category classifiers $\widetilde{W}^U$ to obtain its category label as

$$y = \widetilde{W}^U x_t \tag{10}$$

4. Experiments

4.1. Experimental Settings

We conduct experiments on ImageNet dataset [6], which is a large-scale dataset widely used in zero-shot learning tasks [8, 12, 22, 29]. Following the training and testing split settings used in works [8, 12, 22, 29], we employ three benchmark datasets, i.e., "2-hops", "3-hops" and "All" for evaluation. Note that the classes in the above three datasets and ImageNet 2012 1K classes (1K seen classes) are disjoint. Moreover, we employ the same Hit@k metric [8, 12, 22, 29] which represents the percentage of hitting the ground truth labels in the top k predictions. Meanwhile, we compare the proposed approach with several recent state-of-the-art methods including EXEM [2], GCNZ [29], SGCN [12] and DGP [12].

4.2. Implementation Details

We employ recent ResNeSt-50 [31] model as CNN feature extractor which has been pre-trained on the ImageNet 2012 1K dataset [6]. Following the strategy in previous works [12, 29], we adopt GloVe text model [23] trained on Wikipedia dataset as semantic word features S for all n categories. The GCN module in our model contains two convolutional layers and the weight α is set as 0.2. In constructing K-NN semantic correlation graph G_f, we set K to 2. The number m of negative samples is set to 7 in graph contrastive learning module. For each convolutional layer, we employ Dropout operation and leaky ReLUs with a dropout rate of 0.5 and a negative slope of 0.2 respectively. Our model is trained for 3000 epochs based on Adam [14] algorithm with the learning rate of 0.001 and the weight decay of 0.00005.

4.3. Performance Comparison

We summarize the comparison results on three datasets in Table 1. Here, we can observe that, on all datasets our model performs better than the other related approaches including EXEM [2], GCNZ [29], SGCN [12] and DGP [12]. More specifically, on the "2-hops" dataset, we obtain an obvious improvement than EXEM [2] on top-1 accuracy. Moreover, our model achieves better performance than some other zero-shot learning methods which combine both implicit knowledge (semantic embeddings) and explicit knowledge (knowledge graph) together, i.e., GCNZ [29], SGCN [12] and DGP [12]. This clearly shows the effectiveness and benefit of the proposed graph contrastive learning module to generate more effective and discriminative classifiers for zero-shot classification.

Test Set	Model	Hit@k(%)				
		1	2	5	10	20
2-hops	EXEM [2]	12.5	19.5	32.3	43.7	55.2
	GCNZ [29]	19.8	33.3	53.2	65.4	74.6
	SGCN [12]	26.2	40.4	60.2	71.9	81.0
	DGP [12]	26.6	40.7	60.3	72.3	81.3
	Ours	**28.4**	**43.0**	**62.6**	**74.5**	**82.9**
3-hops	EXEM [2]	3.6	5.9	10.7	16.1	23.1
	GCNZ [29]	4.1	7.5	14.2	20.2	27.7
	SGCN [12]	6.0	10.4	18.9	27.2	36.9
	DGP [12]	6.3	10.7	19.3	27.7	37.7
	Ours	**7.0**	**11.7**	**20.7**	**29.2**	**39.0**
All	EXEM [2]	1.8	2.9	5.3	8.2	12.2
	GCNZ [29]	1.8	3.3	6.3	9.1	12.7
	SGCN [12]	2.8	4.9	9.1	13.5	19.3
	DGP [12]	3.0	5.0	9.3	13.9	19.8
	Ours	**3.3**	**5.6**	**10.1**	**14.7**	**20.5**

Table 1. Top-k performance for different methods on three datasets. Only testing on the unseen classes.

Test Set	Model	Hit@k(%)				
		1	2	5	10	20
2-hops(+1K)	GCNZ [29]	9.7	20.4	42.6	57.0	68.2
	SGCN [12]	**11.9**	**27.0**	50.8	65.1	75.9
	DGP [12]	10.3	26.4	50.3	65.2	76.0
	Ours	7.0	26.8	**52.5**	**67.5**	**77.9**
3-hops(+1K)	GCNZ [29]	2.2	5.1	11.9	18.0	25.6
	SGCN [12]	**3.2**	**7.1**	16.1	24.6	34.6
	DGP [12]	2.9	**7.1**	16.1	24.9	35.1
	Ours	2.0	**7.1**	**17.3**	**26.2**	**36.5**
All(+1K)	GCNZ [29]	1.0	2.3	5.3	8.1	11.7
	SGCN [12]	**1.5**	**3.4**	7.8	12.3	18.2
	DGP [12]	1.4	**3.4**	7.9	12.6	18.7
	Ours	1.0	**3.4**	**8.5**	**13.2**	**19.3**

Table 2. Top-k performance for different methods on three datasets. Testing on both the unseen and seen classes.

Some qualitative comparison results are shown in Figure 2. Intuitively, from the top-5 classification results of each testing image, we can observe that our model obtains better performance on zero-shot classification. Furthermore, following the strategy in works [8, 12, 22, 29], we also conduct the experiments in which the classifiers include both seen class classifiers and unseen class classifiers together for testing. The comparison results of different methods are summarized in Table 2. We can observe that, (1) since the seen class classifiers are added to the classifiers, the performance of all methods drops partly. (2) our model still maintains comparable performance when comparing with GCNZ [29], SGCN [12] and DGP [12]. This further demonstrates the effectiveness of the proposed zero-shot learning

Figure 2. The top-5 classification results on some unseen categories. The correct category is marked as bold.

approach.

4.4. Model Analysis

We conduct ablation study to verify the effectiveness of each component in our proposed method. As shown in Table 3, we implement some variants of the proposed model and report the comparison results. Here, one can note that, (1) the model with WordNet hierarchy knowledge graph G_h performs better than the model with semantic correlation knowledge graph G_f. (2) The model that integrates both G_h and G_f achieves better performance than the model with either G_h or G_f only. This shows that combining both G_h and G_f can capture more complete correlation information among categories. (3) The performance of our model is better than other variants. This demonstrates the capability of our graph contrastive learning module to generate discriminative and effective classifiers.

G_h	G_f	Contrastive Learning	Hit@k(%)				
			1	2	5	10	20
✓	×	×	27.7	42.8	62.5	74.2	82.8
×	✓	×	19.2	31.1	47.8	58.6	67.5
✓	✓	×	27.9	42.8	62.5	74.3	82.8
✓	✓	✓	**28.4**	**43.0**	**62.6**	**74.5**	**82.9**

Table 3. Results of ablation study for the 2-hops dataset. G_h and G_f represent WordNet hierarchy graph and semantic correlation graph, respectively.

5. Conclusion

In this paper, we propose a novel graph contrastive convolutional learning method for zero-shot learning tasks. It explicitly explores multiple relationships among different categories for category classifier learning via dual graph convolutional representation and contrastive learning. The introduced graph contrastive learning module effectively encourages the consistency of category representations from different knowledge graphs while enhancing the discriminability of the predicted category classifiers. Experimental results on several benchmark datasets verify the effectiveness of our zero-shot framework.

6. Acknowledgments

This work was supported in part by the Joint Funds of the National Natural Science Foundation of China under Grant U20B2068, in part by the National Natural Science Foundation of China under Grant 62076004 and in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2020-013.

References

[1] James Atwood and Don Towsley. Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1993–2001, 2016.

[2] Soravit Changpinyo, Wei-Lun Chao, and Fei Sha. Predicting visual exemplars of unseen classes for zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3476–3485, 2017.

[3] Jingjing Chen, Liang-Ming Pan, Zhi-Peng Wei, Xiang Wang, Chong-Wah Ngo, and Seng Tat Chua. Zero-shot ingredient recognition by multi-relational graph convolutional network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10542–10550, 2020.

[4] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1043–1052, 2018.

[5] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision*, pages 48–64, 2014.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[7] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems*, pages 2224–2232, 2015.

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.

[9] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2371–2381, 2021.

[10] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, pages 4116–4126, 2020.

[11] Ming Jin, Yizhen Zheng, Yuan-Fang Li, Chen Gong, Chuan Zhou, and Shirui Pan. Multi-scale contrastive siamese networks for self-supervised graph representation learning. *arXiv preprint arXiv:2105.05682*, 2021.

[12] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11487–11496, 2019.

[13] Youngsung Kim, Jinwoo Shin, Eunho Yang, and Sung Ju Hwang. Few-shot visual reasoning with meta-analogical contrastive learning. In *Advances in Neural Information Processing Systems*, pages 16846–16856, 2020.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[15] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[16] Peike Li, Yunchao Wei, and Yi Yang. Consistent structural relation learning for zero-shot segmentation. In *Advances in Neural Information Processing Systems*, pages 10317–10327, 2020.

[17] Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. Learning a few-shot embedding model with contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8635–8643, 2021.

[18] Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, Xuanyi Dong, and Chengqi Zhang. Isometric propagation network for generalized zero-shot learning. In *International Conference on Learning Representations*, 2021.

[19] Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Attribute propagation network for graph zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4868–4875, 2020.

[20] Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9273–9281, 2020.

[21] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[22] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations*, 2014.

[23] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[24] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

[25] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. Learning to share visual appearance for multiclass object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1481–1488, 2011.

[26] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, D. Christopher Manning, and Y. Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, pages 935–943, 2013.

[27] Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*, 2019.

[28] Chaoqun Wang, Xuejin Chen, Shaobo Min, Xiaoyan Sun, and Houqiang Li. Task-independent knowledge makes for transferable representations for generalized zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2710–2718, 2021.

[29] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6857–6866, 2018.

[30] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, pages 5812–5823, 2020.

[31] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[32] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4166–4174, 2015.

[33] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph contrastive learning with adaptive augmentation. In *Proceedings of the Web Conference 2021*, pages 2069–2080, 2021.